Cross Attention Transformers for Multi-modal Unsupervised Whole-Body PET Anomaly Detection

Ashay Patel https://orcid.org/0000-0003-4212-2578   ashay.patel@kcl.ac.uk

Petru-Danial Tudiosu https://orcid.org/0000-0001-6435-5079   petru.tudosiu@kcl.ac.uk

Walter H.L. Pinaya https://orcid.org/0000-0003-3739-1087   walter.diaz_sanz@kcl.ac.uk

Gary Cook https://orcid.org/0000-0002-8732-8134   gary.cook@kcl.ac.uk

Vicky Goh https://orcid.org/0000-0002-2321-8091   vicky.goh@kcl.ac.uk

Sebastien Ourselin https://orcid.org/0000-0002-5694-5340   sebastien.ourselin@kcl.ac.uk

M. Jorge Cardoso https://orcid.org/0000-0003-1284-2558   m.jorge.cardoso@kcl.ac.uk
King's College London, Becket House, 1 Lambeth Palace Rd, London SE1 7EU

Abstract

Cancer is a highly heterogeneous condition that can occur almost anywhere in the human body. ^{18}F-fluorodeoxyglucose (^{18}F-FDG PET) is a imaging modality commonly used to detect cancer due to its high sensitivity and clear visualisation of the pattern of metabolic activity. Nonetheless, as cancer is highly heterogeneous, it is challenging to train general-purpose discriminative cancer detection models, with data availability and disease complexity often cited as a limiting factor. Unsupervised learning methods, more specifically anomaly detection models, have been suggested as a putative solution. These models learn a healthy representation of tissue and detect cancer by predicting deviations from the healthy norm, which requires models capable of accurately learning long-range interactions between organs, their imaging patterns, and other abstract features with high levels of expressivity. Such characteristics are suitably satisfied by transformers, which have been shown to generate state-of-the-art results in unsupervised anomaly detection by training on normal data. This work expands upon such approaches by introducing multi-modal conditioning of the transformer via cross-attention i.e. supplying anatomical reference information from paired CT images to aid the PET anomaly detection task. Furthermore, we show the importance and impact of codebook sizing within a Vector Quantized Variational Autoencoder, on the ability of the transformer network to fulfill the task of anomaly detection. Using 294 whole-body PET/CT samples containing various cancer types, we show that our anomaly detection method is robust and capable of achieving accurate cancer localization results even in cases where normal training data is unavailable. In addition, we show the efficacy of this approach on out-of-sample data showcasing the generalizability of this approach even with limited training data. Lastly, we propose to combine model uncertainty with a new kernel density estimation approach, and show that it provides clinically and statistically significant improvements in accuracy and robustness, when compared to the classic residual-based anomaly maps. Overall, a superior performance is demonstrated against leading state-of-the-art alternatives, drawing attention to the potential of these approaches.

Keywords: Transformers, Unsupervised Anomaly Detection, Cross Attention, Multi-modal, Vector Quantized Variational Autoencoder, Whole-Body, Kernel Density Estimation

1. Introduction

Cancer is a disease affecting approximately one in two people over their lifetime (Ahmad et al., 2015). In 2020 alone, over 19 million new cases were reported worldwide, a figure expected to rise by over 50% by 2040 (Sung et al., 2021). Although preventative measures can be taken through dietary and lifestyle changes, often the first real line of defense is through early diagnosis of which medical imaging plays a key role. Amongst imaging modalities, ^{18}F-fluorodeoxyglucose Positron Emission Tomography (^{18}F-FDG PET) has one of the highest detection rates for cancer (Liu et al., 2017; Endo et al., 2006). Through enabling the visualization of the glycolytic pathway, the efficacy of ^{18}F-FDG PET is related to the high metabolic rates of cancer cells (Almuhaideb et al.). As such PET may enhance cancer staging, treatment planning, and the evaluation of patient responses to treatment (Kim et al., 2015). PET is almost always coupled with CT, and more recently hybrid PET/MRI scanners have been introduced into clinical practice. This allows anatomical localization of ^{18}F-FDG uptake as well as attenuation correction of the PET signal. From an unsupervised anomaly detection approach, using CT as an anatomical point of reference by combining modalities can further enhance PET interpretability, exemplified in Fig 1.
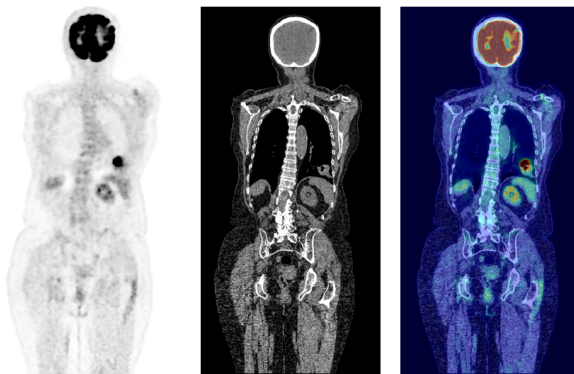

Figure 1: ^{18}F-FDG PET/CT Imaging. PET (left), CT (centre), combined CT-PET (right). High intensity area in lower chest seen in PET and CT-PET scan shows cancer. Note the non-cancer related high PET signal in the brain and kidney regions.

In clinical practice PET images are usually read in a qualitative manner, across a range of clinically relevant tasks from staging, treatment planning, and surgical or therapy intervention planning. Sensitivities can range as much as 35% for PET imaging depending on the nuclear medicine/PET physician and cancer type (Newman-Toker et al., 2021). This can be an issue in the case of metastatic cancer where small lesions can be overlooked (Perani et al., 2014). Considering these shortfalls, there is significant motivation for developing accurate automated detection methods, a major topic of interest in medical imaging research.

1.1 Related Work

Quantitative imaging analysis is an approach aimed to tackle this problem through segregating normal and pathological findings by finding optimal thresholds. This approach

can be done via regional or voxel-wise analysis. For regional analysis, uptake is compared to the same regional uptake found in a healthy control population. This analysis however requires extensive prior knowledge of the subject and control population in order to select the most valid atlases and relevant discriminant regions greatly limiting this approach's efficacy (Signorini et al., 1999). In classic voxel-wise analysis, a PET image is often registered to a normal standardised group space to compare voxel-wise differences in uptake. Such approaches have been implemented in Neurostat and NeuroGam (Drzezga et al., 2005; Renard et al., 2013). Similar implementations have been carried out to generate Z-score maps to demonstrate the degree of abnormality between the healthy model and individual subjects. This approach implemented in Burgos et al. (2021) developed patient-specific models through using a healthy control atlas database. However, these tools have primarily been developed for brain imaging with limited further use outside this area, in addition the approach has sub-optimal assumptions about the statistical occurrence of abnormalities, i.e. noise models and parametric distributions.

Breakthroughs in recent years have primarily showcased the efficacy of deep learning models for anomaly detection in medical data. Given the limitations of previous approaches on whole-body data, the use of deep learning shows promise in the task of anomaly detection.

Unsupervised methods have become an increasingly prominent field in recent years for automatic anomaly detection by eliminating the necessity of acquiring accurately labelled data and showing a strong ability to generalise to unseen anomalies (Chen et al., 2020; Baur et al., 2020). These methods mainly rely on creating generative models trained solely on non-anomalous data. Then during inference, anomalies are defined as deviations from the defined model of normality as learnt during training. However, their efficacy is often limited by the requirement of uncontaminated (e.g., non-anomalous) data with minimal anomalies present during training.

One particular model used for unsupervised anomaly detection is the family of generative adversarial networks (GANs). GANs are able to generate data without explicitly modelling the probability density function of the underlying data (Yi et al., 2019). The overall architecture consists of two sub-networks, namely a discriminator and generator. During training, the generator is tasked with generating samples from a given input (usually gaussian noise). The objective is to generate samples that the discriminator cannot differentiate from real. The use of GANs has many applications in medical imaging, including reconstruction (Quan et al., 2018), denoising (Armanious et al., 2018; Kang et al., 2018) and cross modality translation (Armanious et al., 2018; Bi et al., 2017; Ben-Cohen et al., 2017; Armanious et al., 2019). GANs also show promise in the field of anomaly detection, as demonstrated in (Schlegl et al., 2017; Alex et al., 2017; Sun et al., 2018). The approach taken by Sun et al. (2018) for brain MRI anomaly detection used CycleGan, trained to generate healthy-looking scans from anomalous ones. The approach however was shown to be imperfect brought down to textual differences and ununiform intensities during the reconstruction of abnormalities in addition to the instabilities inherent in training GANs models, making them more prone to model collapse (Kodali et al., 2017). As such im-

plementing a GAN architecture poses a greater challenge for the task of 3D image synthesis.

The most prevalent competing family of networks for unsupervised anomaly detection is the Autoencoder (AE). AEs are models that are made of two main architectures, an encoder and decoder. The encoder maps the input image into a lower dimension manifold z, where the decoder will then reconstruct the original image from said manifold. The driving characteristic of AEs that make them suitable for anomaly detection is the low dimensional manifold bottleneck; for an AE sufficiently constrained and trained on healthy data only, the model will struggle to generalise when faced with unseen anomalies and will have large reconstruction errors associated at these locations (Baur et al., 2020). Such approaches have been used in various applications, such as brain tumour detection in MRI and CT (Atlason et al., 2018; Baur et al., 2019; Sato et al., 2018). However, a common problem with conventional AEs is the lack of regularisation of the latent space, reducing its efficacy for anomaly detection (Chen and Konukoglu, 2018). Furthermore, AEs are simply a learnt function whose purpose is to compress data and minimise reconstruction loss. As such, an AE can easily overfit data unless explicitly regularised and can render irregular reconstructions on unseen data. Improving on these AE limitations, spatial VAEs have been proposed Baur et al. (2020). Here, the healthy data manifold is obtained by constraining the latent space to conform to a given distribution. A comparative study has shown that VAEs demonstrated improved performance as an alternative to AEs (Baur et al., 2020). The same study also revealed an improvement against adversarial-based strategies like f-AnoGAN (Schlegl et al., 2017). Even with such improvements however, a particular weakness, as explored in Makhzani et al. (2015), is that the KL-divergence prompts the posterior to incorporate the mode of the prior, not necessarily the entire distribution. This can result in an oversimplified prior brought about by the KL-divergence term resulting in over-regularization. Furthermore, the objective of the VAE can result in trivial solutions that decouple the input from the model's latent space resulting in posterior collapse (Chen et al., 2016). This approach is further limited by low fidelity reconstructions and unwanted reconstructions of unseen pathologies suggesting a shortfall in the model itself.

To overcome some of these issues, an approach for unsupervised anomaly detection was presented utilising autoregressive models coupled with a superior autoencoder model, namely the vector-quantised variational autoencoder (VQ-VAE) (van den Oord et al., 2017; Marimont and Tarroni, 2020). Transformers, currently state-of-the-art networks in the language modelling domain (Vaswani et al., 2017; Radford and Narasimhan, 2018), use attention mechanisms to learn contextual dependencies regardless of location, allowing the model to learn long-distance relationships to capture the sequential nature of input sequences. This general approach can be generalised to any sequential data, and many breakthroughs have seen the application of transformers in computer vision tasks from image classification to image and video synthesis (Chen et al., 2020; Child et al., 2019; Yan et al., 2021). Although having showcased state-of-the-art performance in unsupervised anomaly detection tasks for medical imaging data (Pinaya et al., 2021), these methods still rely heavily on purely normal data for model training. To date little research has been carried out using unlabelled (image-wise or pixel-wise) training data that contain anomalies. The work in Zhang and Zhuang (2022) and Zuluaga et al. (2011) proposes methods that make use of anomalous

training data, but even so for this method to work, an initial portion of labelled normal training data is required to successfully make use of the unlabelled training portion. To the best of our knowledge, no prior research exists using unsupervised methods to accurately localise anomalies while using training data containing such anomalies with no prior knowledge over any samples whether they they contain anomalies or not. This task itself is of importance given the nature of whole-body PET. It is often difficult or unethical to obtain healthy datasets of certain medical imaging modalities as some images are only acquired with prior suspicion of disease.

1.2 Contributions

To address these problems and shortfalls of existing state-of-the-art techniques, we propose a method for unsupervised anomaly detection and segmentation using multi-model imaging via transformers with cross attention. Leveraging off previous work combining the use of VQ-VAE models with transformers we propose, deploy and evaluate the following contributions in our work:

- We show the added benefit of the use of multi-modal imaging for unsupervised anomaly detection, achieved through the use of transformers with cross attention

- We highlight the importance of optimal choices in the VQ-VAE codebook architecture, beyond accurate reconstruction performance, and its effect further downstream on the transformer's performance for anomaly detection

- We introduce an improved alternative to commonly used residual-based anomaly maps via a kernel density estimation approach

- We supplement this kernel density estimation approach with an extensive study of kernel choices and a selection of regularisation parameters.

- We carry out our training on data where healthy samples are unattainable and still show high detection rates during testing where other models fail

This work is an extended version of a conference workshop paper presented at DGM4MICCAI (Patel et al., 2022). The extensions to the conference paper involve a more substantial literature review in addition to a larger dataset that has yielded a greater number of training samples and a higher number of samples for testing on unseen cases. Additionally a comprehensive ablation study showcasing the importance of codebook sizing in the Vector Quantized-Variational Autoencoder is explored that can highlight its affect on the performance of the transformer model for anomaly detection. A deeper exploration into the kernel density estimation approach is also carried out showcasing the difference in performance of varying kernel choices in addition to varying regularisation parameters for this methodology. Finally, as further validation of our methods, we carry out an additional set of testing on a fully out-of-sample testing dataset with varying cancer cases from a different source to that of the original training and testing data.

176

2. Proposed Method

The principal components behind the proposed whole-body anomaly detection model relies on using transformer models and auto-encoders to model 3D whole-body ^{18}F-FDG PET scans. Although all training data contain anomalies, the spatial distribution of anomalies across samples will result in such anomalies being unlikely, thus appearing at the likelihood tail-end of the learnt distribution. In order to use transformer models, images need to be expressed as a sequence of values, ideally categorical. As it is not computationally feasible to do this using voxel values, a compact quantized (discrete) latent space is used as input for the transformer via a VQ-GAN model as named and proposed in Esser et al. (2020) (a VQ-VAE van den Oord et al. (2017) with an adversarial component).

2.1 VQ-GAN

The original VQ-VAE model (van den Oord et al., 2017) is an autoencoder that learns discrete latent representations of images. The model comprises of three principal modules: the encoder that maps a given sample $X \in \mathbb{R}^{H \times W \times D}$ onto a latent embedding space $e \in \mathbb{R}^{h \times w \times d \times n_z}$ where n_z is the dimension of each latent vector e_i. After the encoder network projects the image X to its latent representation $z_e(X)$, the discrete latent variables z are generated by a nearest neighbour look-up to the shared embedding space e to generate $e_m, m \in 1, ... M$ where M is the vocabulary size. For simplicity we can refer to a single random variable as z to represent a single discrete latent variable, given the input this can represent a 1D, 2D or 3D latent feature space. In this case given we are dealing with 3D medical data z corresponds to a 3D feature space. During training the codebook is learnt jointly with model parameters. The posterior distribution can then be given as a categorical one defined as:

$$q(z = m|X) = \begin{cases} 1 \ for \ m = \text{argmin}_j \|z_e(X) - e_j\|_2 \\ 0 \ otherwise \end{cases} \tag{1}$$

Here $z_e(X)$ is the output of the encoder giving us our encoded feature vectors that are matched to e_j, a learnt codebook vector in the shared embedding space. The discrete latent space representation is thus a sequence of indexes for each code from the codebook. The final portion of the network is the decoder, which reconstructs the original observation from the quantized latent space. The total loss for the VQ-VAE is then given as:

$$L_{VQVAE} = \left\|(X - \hat{X})\right\|_2^2 + \left\||STFT(X)| - \left|STFT(\hat{X})\right|\right\|_2^2 + \|sg[z_e(X)] - e\|_2^2 + \beta \|z_e(X) - sg[e]]\|_2^2 \tag{2}$$

where $\hat{X}$ is the output from the decoder and sg is a stop gradient operator to stop gradients from flowing back into their argument. The loss function for the VQ-VAE makes use of a spectral loss (Dhariwal et al., 2020) that is, it includes a component based on the magnitude of the Fourier transform of the original and reconstructed image. From equation 2 the first term is the pixel loss, the second term is the spectral loss between the original and reconstruction where SFTF stands for the short time Fourier transform. The third term is the commitment cost used to ensure the encoder commits to the codebook. The final term is to move the codebook embedding vectors towards the output from the encoder. For this term, we replace this and use the exponential moving average updates for the codebook

(van den Oord et al., 2017). During training, a β of 0.25 was used.

As autoencoders often have limited fidelity reconstructions (Dumoulin et al., 2016), an adversarial loss is added to the VQ-VAE network to form a VQ-GAN, as proposed and named in Esser et al. (2020). When implementing the VQ-GAN network, however, due to instabilities associated with adversarial networks, the loss function is further expanded to include a perceptual loss (Takaki et al., 2018) that helps preserve spatial consistency by using the LPIPS library (Zhang et al., 2018). The perceptual loss applied via the LPIPS library makes use of 2D images and as such has to be applied over slices of the original and reconstructed sample. To improve efficiency during training the loss is randomly applied to 50% of slices across each plane. The adversarial loss used is of the following form:

$$\mathcal{L}_{adv} = \mathcal{L}_{LSGAN}(Dis) + \mathcal{L}_{LSGAN}(E, Quant, D) \tag{3}$$

Where Dis is the discriminator and $E, Quant, D$ is the VQ-VAE Encoder, Quantizer and Decoder Respectively. This is based on the Patch-GAN model (Isola et al., 2017) as per (Esser et al., 2020) and paired with the LS-GAN loss (Mao et al., 2017) providing a more stable and reproducible behaviour, denoted as:

$$\min_{Dis} L_{\text{LSGAN}}(Dis) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\left[(Dis(X)-1)^2\right]$$
$$+ \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\mathbf{data}}(\mathbf{x})}\left[(Dis(D(Quant(E(X)))))^2\right]$$
$$\min_{Enc,Quant,Dec} L_{\text{LSGAN}}(E, Quant, D) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{z})}\left[(Dis(D(Quant(E(X))))-1)^2\right]$$

2.2 Transformer

Once a VQ-GAN model is trained, we now are required to train a generative model on the discrete latent representation. For this we use a transformer. Transformer models rely on attention mechanisms to capture the relationship between inputs regardless of the distance or positioning relative to each other. The self-attention mechanism is best described as a mapping of intermediate representations of three position-wise linear layers onto three representations denoted by the Value (V), key (K) and query (Q) (Vaswani et al., 2017). With d_k denoting the dimension of the key vectors, the attention mechanism is calculated as:

$$Attn(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

The multi-head attention aspect of this transformer network is then several attention layers run in parallel with their outputs concatenated and fed through a linear layer. This process, however, relies on the inner product between elements and, as such, network sizing scales quadratically with sequence length. Given this limitation, achieving full attention with large medical data, even after the VQ-GAN encoding, comes at too high a computational cost. To circumvent this issue, many efficient transformer approximations have been proposed (Tay et al., 2020; Choromanski et al., 2020). In this study, a Performer

model is used; the Performer uses the FAVOR+ algorithm (Choromanski et al., 2020), which proposes a linear generalized attention that offers a scalable estimate of the attention mechanism. Using such a model, we can apply transformer-like models to much longer sequence lengths associated with whole-body data. In order to learn from the discrete latent representations, we require the discretised latent space z_q to take the form of a 1D sequence s using some arbitrary ordering. The transformer is then used to model s by minimizing the conditional distribution $p(s_i) = p(s_i \mid s_{<i})$ where i is the ith element of s.

2.3 Anomaly Detection

To perform the baseline anomaly detection model on unseen data, as proposed by Pinaya et al. (2021), first, we obtain the discrete latent representation of a test image using the VQ-GAN model. Next, the latent representation z_q is reshaped using a 3D raster scan into a 1D sequence s where the trained Performer model is used to obtain likelihoods for each latent variable. At each position in the sequence the trained transformer will give the learnt likelihood of each possible token appearing at every point in the sequence. In doing so we can highlight low likelihood (or anomalous tokens) as $p(s_i) = p(s_i \mid s_{<i}) < t$, (where t is a threshold determined empirically using a validation dataset; t = 0.025 was found to be optimal). This generates a binary resampling mask that indicates which tokens in the latent sequence are anomalous i.e. below the threshold. Using the resampling mask, the anomalous latent variables are removed and replaced in the sequence with non-anomalous tokens by resampling from the transformer. This approach replaces anomalous latent variables with those that are more likely to belong to a healthy distribution, as such "healing" the considered anomalous latent space. Using the non-anomalous latent space, the VQ-GAN model reconstructs the original image X as a non-anomalous reconstruction C_r. Finally, a voxel-wise residual map can be calculated as $X{-}X_r$ with final segmentations calculated by thresholding the residual values. As areas of interest in PET occur as elevated uptake, residual maps are filtered to only highlight positive residuals.

2.4 CT Conditioning

There are often times when more information can be useful for inference. This can be in the imaging domain through multiple resolutions (Chen et al., 2021) or multiple modalities/spectrums (Mohla et al., 2020). It is for these tasks where cross-attention can prove beneficial. From a clinical point of view, whole-body PET scans are acquired in conjunction with CT, or less frequently MRI data for attenuation correction purposes in addition to providing an anatomical reference. Additionally, it can be observed that areas of high uptake are not always associated with pathological findings i.e. high uptake may reflect physiological uptake, e.g. within the brain and heart. Additionally areas where radiotracer may collect like the kidney and bladder can also show high uptake patterns. Acknowledging these areas of high physiological uptake by recognition of the organ location with respect to a whole 3D scan visible may seem obvious to the human eye, however this may not be the case using the transformer approach. For this work images are encoded to a discrete latent space and then rasterized into a 1D sequence. During training and inference the model works in an autoregressive manor i.e. only prior tokens in the sequence can be viewed. As such when looking at a specific token (and only the prior tokens in the sequence), it

may be hard for the model to determine the exact anatomical point within the whole body that that specific token represents, without further context relating to the whole body. As such the anatomical reference provided from CT data is beneficial. This leads to one of the main contributions of the work, namely anomaly detection incorporating CT data. This process works by generating a separate VQ-GAN model to reconstruct the PET-registered CT data. Then, CT and PET data are encoded and ordered into a 1D sequence using the same rasterization process, such that CT and PET latent tokens are spatially aligned. The transformer network is then adapted to include cross-attention layers (Gheini et al., 2021) that feed in the embedded CT sequence after each self-attention layer. At each point in the PET sequence, the network has a full view of the CT data helping as a structural reference. In doing so, the problem of determining the codebook index at a given position i becomes $p(s_i) = p(s_i \mid s_{<i}, c)$, where c is the CT latent sequence.
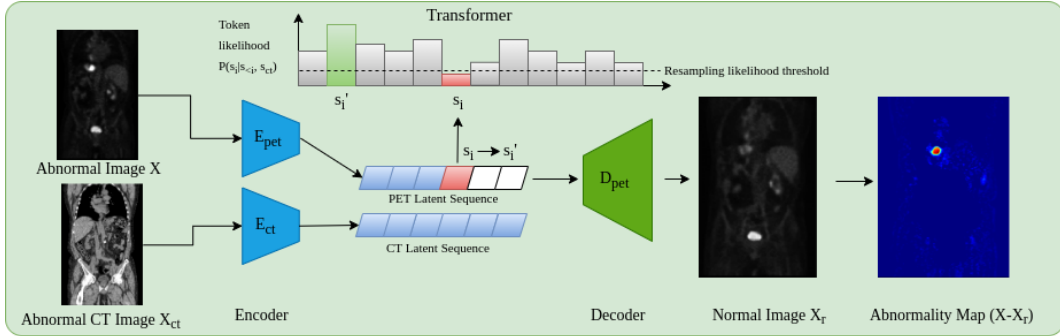

Figure 2: Anomaly Detection Pipeline – PET image X is encoded along with CT image X_{ct}. Tokens form the encoded PET image are then sampled from the transformer by obtaining their likelihood with respect to prior tokens in the sequence and all CT tokens. Tokens below a given threshold are resampled from a multinomial distribution, derived from likelihood outputs from the transformer for all tokens at a given position in the sequence. This yields a non-anomalous latent space which is decoded to give X_r.

To add cross attention to the transformer architecture, we add a cross attention layer after each self-attention layer in the transformer architecture. Still using the same attention mechanism, the cross attention calculation is then given as:

$$Attn(Q_s, K_c, V_c) = softmax\left(\frac{Q_s K_c^T}{\sqrt{d_k}}\right) V_c \qquad (5)$$

Where Q_s is the output from the prior self-attention layer, and K_c and V_c are the Key and Query vectors derived from the embedded conditioning CT sequence. The architecture of an entire transformer layer with self-attention and cross-attention can be visualised as in Fig. 3:

This approach for anomaly detection, as visualised in Fig. 2, adds robustness to the anomaly detection framework by providing meaningful context in areas of greater variability in uptake that can be explained by the anatomical information within CT.
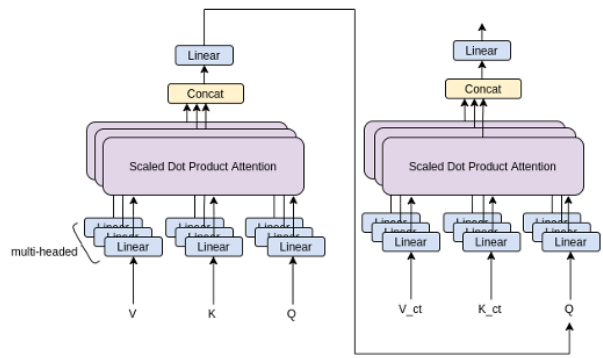
Figure 3: Network architecture showcasing layers in transformer with multi-headed self-attention (left) and corresponding multi-headed cross-attention (right) run in series

2.5 Kernel Density Estimation

A drawback of the baseline anomaly detection method described in section 2.3 is that the residual image uses an explicit image-wide threshold to generate a segmentation map. The resulting segmentation can often be noisy due to discrepancies between the reconstructed image and the original, for example, between borders of high intensity. Additionally, anomalies can occur at different intensities, meaning a blanket threshold is inappropriate. A possible solution is to implement Z-score anomaly maps as used in similar anomaly detection work (Burgos et al., 2021). For this work, this can be achieved by introducing stochasticity within the model. As the distribution of uptake patterns is often multi-modal, due to its relationship to base metabolic rate and procedure-related variations (eg injected tracer amount and time since injection), the optimality of the Z-score's Gaussian-error assumption should be questioned and relaxed. Empirical evidence obtained by exploring the data and by sampling from the transformer itself highlights that the error is indeed non-Gaussian even in non-anomalous regions. For example, in the heart, bi-modal (even multi-modal) error distributions are observed. To remedy this, we propose to use a non-parametric approach using kernel density estimation (KDE) (Parzen, 1962). To do this, we sample from the model by introducing a dropout layer in the VQ-GAN decoder. Additionally, we achieve further stochasticity by replacing unlikely tokens with ones drawn from a multinomial distribution, derived from the likelihoods output from the transformer for each token at a given position in the sequence. During inference, by sampling multiple times, we generate multiple normal latent representations for a single image, which are then decoded multiple times with dropout to generate multiple non-anomalous reconstructions of a sample, at which point a KDE is fit independently at each voxel position to generate an estimate of the probability density function f for the intensities at a specific point across reconstructions. Letting $(x_1, ..., x_n)$ be the intensity for a voxel position across reconstructions, we can generate an estimation for the shape of the density function f for voxel x as:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{6}$$

Here, K is a given kernel shape and h is a smoothing bandwidth calculated via the Silverman method (Silverman, 2018) as:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} + \epsilon \tag{7}$$

with $\hat{\sigma}$ representing the standard deviation at the given voxel position across n reconstructions and ϵ is a scalar regularisation parameter determined empirically on the validation dataset used to address areas of low variance across reconstructions. We can then score voxels from that estimated density function at the intensity of the real image, at the voxel level, to generate a log-likelihood for that intensity, generating the anomaly map. The KDE approach can be further explored via varying kernels and regularisation parameters. As part of this study, we evaluate the use of a series of different Kernels to calculate the estimation of the probability density function f in addition to varying regularisation values for ϵ.

2.6 Clinically Consistent PET Segmentations

For whole-body PET, due to the large system Point Spread Function (PSF) of observed uptake, the contours of an anomaly can be hard to define. Given this there are a number of methods used clinically to define the boundaries of anomalies. This can range from absolute thresholding over the entire image, to per anomaly thresholding via percentages of the maximum uptake of individual anomalies, to more involved and computationally expensive algorithms (Berthon et al., 2017; Hatt et al., 2017). In general there is no universal standard for defining the anomaly boarders and as such we use the method most recognised as the clinical standard in the UK as advised by our clinical experts. The method used makes use of percentage thresholding for individual anomalies. This defines boundaries of an anomaly as connecting voxels with intensities above 40% of the maximum intensity of a specific anomaly and has shown to generate optimal performance with limited computation (Berthon et al., 2017). To conform to this standard, we apply a final post-processing step of growing all initial segmentations to satisfy this criteria.

3. Data

For the training, validation and testing of the methods described above, a combination of two datasets was used to overcome the limitations of limited data for training purposes and the lack of validity for the proposed methods during testing in the case of limited testing samples. Furthermore, to improve the efficacy of the approach to unseen out-of-distribution data, the approach will likely be supplemented with training from data that have undergone different acquisition protocols and varying voxel dimensionality. All data were rigidly registered to a group-wise space with a field of view from the neck down to the upper thigh region. The initial step of the group-wise registration was to register all samples to a given sample in the private dataset meaning all datasets were converted to the same voxel spacing, after which, all CT images were registered to their paired PET image using

rigid transformations. All processed images had a final dimension of $168 \times 208 \times 216$. CT data were further preprocessed to remove the bed from the images and have voxel intensities clipped to showcase the soft-tissue window only.

3.1 Private dataset

The private dataset used consists of 83 co-registered ^{18}F-FDG PET/CT images acquired using a GE Medical Systems scanner. The original PET images had voxel dimensions of $2.7 \times 2.7 \times 3.3mm$, whilst the CT images had dimensions of $1.4 \times 1.4 \times 3.3mm$. The dataset comprised a variety of subjects showcasing various primary cancers in various locations across the body and metastases located in numerous further locations. Out of the 83 samples, 60 were used for training whilst the remaining 23 were split, with 12 used for validation and 11 used for testing. From the training data, a total of 49 cases showcased some form of cancer in the scan.

3.2 NSCLC Radiogenomics Dataset

The NSCLC Radiogenomics dataset comprises 211 co-registered ^{18}F-FDG PET/CT samples presenting Non-small cell lung cancer cases (Bakr et al., 2017, 2018; Gevaert et al., 2012; Clark et al., 2013a). The acquisition protocol ranges by candidate using both GE Medical Systems and Siemens Scanners. The original PET images had voxel dimensions of $3.6 \times 3.6 \times 3.3mm$, while the CT images had dimensions of $1.4 \times 1.4 \times 3.3mm$. Out of the 211 samples, 160 were used for training, with 26 used for validation and 25 used for testing. From this training data, all cases contained some form of lung cancer along with potential metastases.

From combining the private dataset and NSCLC Radiogenomics dataset, during training, a total of 220 samples were used, of which 209 samples had some form of anomaly present in the scan. Furthermore 38 samples were used for validation to tune hyperparameters and run ablation studies on model parameters like codebook sizing and Kernel types. The remaining 36 samples were left as a hold-out set for testing on the final models chosen after all ablation studies were carried out.

3.3 AutoPET Dataset

For a comparison of our methods and baselines when trained on fully normal data we leverage the autoPET dataset - a 3D whole-body Positron Emission Tomography (PET) dataset (Gatidis et al., 2022; Clark et al., 2013b). Generally speaking in PET imaging, scans void of any forms of anomalies are hard to come by, as often scans are taken with a strong prior suspicion of a pathology. There are some cases where this may not be true, that includes scans following treatment, which is where the normal samples for this dataset are obtained. This dataset consists of 1014 PET scans with 430 non-anomalous scans, with the remaining containing some form of lung cancer, lymphoma, or melanoma. From this dataset we generate a separate training dataset with normal cases only. We use all 430 healthy scans to form the training data. For this work the same validation set consisting of the private data and NSCLC radiogenomics data is used to tune the model and anomaly

detection hyperparameters, as testing is carried out on their testing set as well for the most fair comparison. The original PET images have voxel dimensions of $2.036 \times 2.036 \times 3mm$. All scans are aligned to the same space as the training data via affine transformations, i.e. a field of view from the neck region to upper thigh region is used out of the whole body images.

3.4 Clinical Proteomic Tumor Analysis Consortium - CPTAC

The CPTAC dataset comprises a combination of several individual CPTAC studies that cover numerous cancers from lung, ovarian, pancreas and skin cancer (CPTAC, 2018; Clark et al., 2013a). A total of 14 co-registered whole-body ^{18}F-FDG PET/CT images are used to demonstrate the ability of the proposed methods on fully out-of-sample data that is not used during training or validation. The data is similarly acquired using a GE Medical Systems Scanner. The original PET images had voxel dimensions of $3.6 \times 3.6 \times 3.3mm$, whilst the CT images had dimensions of $1.4 \times 1.4 \times 3.3mm$. All scans are registered to the same space as the training data.

4. Experiments and Results

4.1 VQ-GAN Training and architecture details

The training details and architecture for the VQ-GAN (besides codebook sizing) remains the same through all experiments run in this study. The architecture used for the VQ-GAN model uses an encoder consisting of three strided convolutional layers with stride 2 and kernel size 4. Each convolutional layer is then followed by a ReLU activation and 3 residual blocks (consisting of a 3x3x3 conv, ReLU, 1x1x1 conv, ReLU). The decoder similarly has 3 residual blocks, each followed by a transposed convolutional layer with stride 2 and kernel size 4. Finally, before the last transposed convolutional layer, a Dropout layer with a probability of 0.05 is added. Further hyperparameters include the use of β equal to 0.25, as stated in equation 2. This value is taken from the original implementation of the VQ-VAE as stated in van den Oord et al. (2017). Through the ablation study exploring the effect of ranging codebook sizes, several codebook dimensions were explored, consisting of atomic elements from 64-2048 with lengths ranging from 32-256. However, changes in the codebook parameters had no change to the encoder and decoder architecture. To train the VQ-GANs, we used an ADAM optimiser (Kingma and Ba, 2014) with a learning rate of 1e-4 and an exponential learning rate decay with a gamma of 0.9999. Additionally, the discriminator network had a learning rate of 5e-4. Training data was augmented using elastic deformations, Gaussian noise, intensity shifts, contrast adjustments and gaussian blur. The model was trained over 1000 epochs with a batch size of 3. Further details on model complexity and the number of model parameters can be seen in table 4 in Appendix A.

4.2 Transformer Training and architecture details

The performer in all experiments used corresponds to a decoder transformer architecture with 16 layers, each with 8 heads and an embedding size of 256. To train the performer network, we used an ADAM optimiser (Kingma and Ba, 2014) with a learning rate of 1e-3

and an exponential learning rate decay with a gamma of 0.9999. The loss function used for training was cross-entropy, given the discrete nature of the latent sequence codes. To obtain the input data for training the transformer network, the trained VQ-GAN model encoded the training data into the discrete latent codes, which were used as inputs for the transformer. To avoid overfitting, the original training dataset was augmented 4 times and then encoded using elastic deformations, Gaussian noise, intensity shifts, contrast adjustments and gaussian blur to increase the number of samples for training. The model was then trained over 120 epochs with a batch size of 1. For the AutoPET dataset the models were trained for 80 epochs with a batch size of 1. Further details on model complexity and the number of model parameters for the Performer model with and without cross attention can be seen in table 4 in Appendix A.

4.3 Experiment 1: PET-only ablation study

The first study explores the effect of codebook size on the anomaly detection pipeline. To do this, we experiment with codebooks ranging vocabulary sizes from 64-2048 and vector dimensions from 32-256. From this range, 24 different VQ-GAN networks of ranging codebooks were trained, of which a further 24 transformer networks were trained using their respective VQ-GAN model using PET data only. Using the trained networks, anomaly detection was run on the validation dataset where anomalies were located via residual maps between the original and reconstructed images. We measure our models' performance using the best achievable DICE score, which serves as a theoretical upper-bound to the models' segmentation performance. This work makes use of the anomalous training dataset. The results can be visualised in Fig. 4
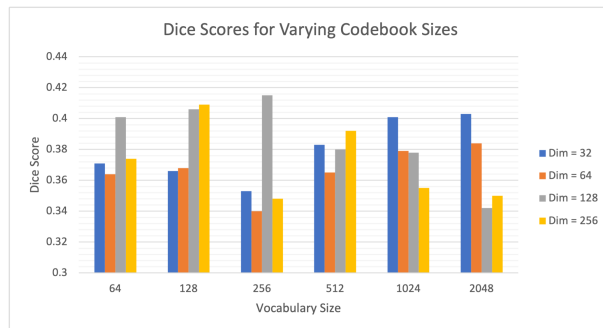

Figure 4: Best achievable DICE-score ($\lceil DICE \rceil$) on validation data for anomaly detection with PET data only and varying codebook vocabulary sizes and latent code dimensions

We can see a large range (0.332-0.415) in DICE scores from adjusting the codebook parameters highlighting the importance of choosing an optimal codebook size for anomaly detection. Furthermore, a general trend can be seen from the results showcasing that codebooks with a smaller vocabulary size generally perform better with large dimensions, whereas when the vocabulary size increases, the network generally performs better with

smaller latent vector dimensions. From the results, however, a codebook with a vocabulary size of 256 with dimension 128 showed to generate the best results.

4.4 Experiment 2: CT Conditioning Ablation Study

Continuing with showcasing the importance of the codebook size for anomaly detection, we further explore ranging codebook sizes with respect to the conditioning sequence from the CT data with results showcased in Table 1. For the experiment, we generate 4 different VQ-GAN models to encode CT images with ranging codebook sizes, chosen and conforming to the results seen in Experiment 1, i.e., small vocabulary size with large dimensions and large vocabulary sizes with smaller dimensions. For each of these experiments, we use a PET VQ-GAN network with the optimal codebook size as derived from Experiment 1, i.e., vocabulary size of 256 with dimension 128. As with Experiment 1, we train the models on the anomalous training dataset and run anomaly detection on the validation set to generate DICE scores using residual maps. As in Experiment 1, a vocabulary size of 256 and dimension of 128 was found to produce the best dice score.

Table 1: Anomaly detection results on whole-body PET validation data for varying codebook sizes. The performance is measured with best achievable DICE-score ($\lceil DICE \rceil$) and AUPRC on the test set.

Vocabulary Size	Dimensions	$\lceil DICE \rceil$
64	256	0.424
256	128	0.458
512	64	0.432
1024	32	0.443

4.5 Experiment 3: KDE Ablation Study

When implementing the kernel density estimation approach, the shape of the probability density estimation relies on 2 factors, the shape of the kernel used and the bandwidth. Given this, we explore varying kernels in addition to a range of ϵ values (the regularisation term as seen in equation 7). In total, six different kernels are used (i.e. gaussian, top-hat, Epanechikov, exponential, linear and cosine), of which their shape can be seen below in Fig. 5.

Furthermore, three ϵ values of 0.025, 0.05 and 0.1 were experimented with. From experiments 1 and 2, we chose the best vocabulary sizes for the PET VQ-GAN and CT VQ-GAN, which is a vocabulary size of 256 and dimension of 128 for both models. As we use the kernel density estimation approach, not residual maps, we rely on multiple reconstructions achieved through resampling from the transformer 60 times, with each latent sequence decoded with dropout four times each. As with the previous experiments score, we measure our models' performance using the best achievable DICE score based on the KDE anomaly maps.

From the results, there are two clear standout kernels that showcase the best performance, namely the gaussian and exponential kernel. The shape of the kernels implies that
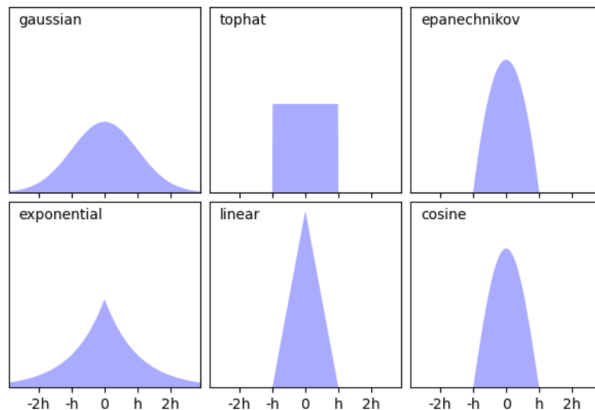
Figure 5: Different Kernel shapes with respect to bandwidth (h) explored in KDE ablation study (Scikit-Learn)
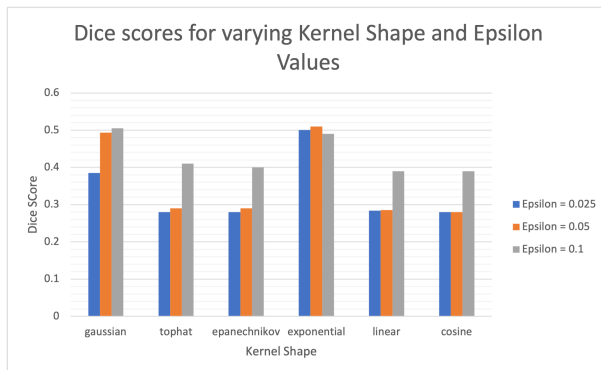

Figure 6: Best achievable DICE-score ($\lceil DICE \rceil$) on validation data for Kernel Density Estimation approach with varying Kernel shapes and ϵ values

kernels with heavier tails seem to be more favorable for the task. Additionally, we generally see higher ϵ values, and therefore higher bandwidths and longer tails, give better performances, with large reductions in performance associated with the smallest ϵ values due to under regularisation and noisy KDE maps. Furthermore, we see the exponential kernel produces the best results in addition to the least variance associated with ranging ϵ values indicating less reliance on an optimal choice of ϵ and perhaps a more robust approach when facing unseen data.

From the series of experiments optimized on the validation data, a final model is generated using a PET and CT VQ-GAN model with a codebook vocabulary size of 256 and dimension 128. Furthermore, on the kernel density estimation side, final parameters are set to calculate probability density estimates using an exponential kernel with an ϵ value of 0.05.

4.6 Testing

After obtaining the optimal models from each hyperparameter study, we test the proposed models on the hold-out testing data as a final ablation study to showcase the added contribution of each proposed method. We compare our results to that of a Dense, Spatial AE and Dense VAE model (Baur et al., 2020) and an SSIM AE (Bergmann et al., 2019). These models are trained on PET only, however we also implement them with 2 channel inputs with PET and CT to showcase the effect of CT conditioning on the baseline models in comparison to the effects of conditioning for our proposed method. We additionally implement F-AnoGan for a further baseline comparison (Schlegl et al., 2017). Testing is carried out on 36 samples of which the best achievable DICE score is calculated in addition to the area under the precision-recall curve (AUPRC) as a suitable measure for segmentation performance under class imbalance. These results can be seen tabulated in Table 2 and visualised in Fig. 7 qualitatively showcasing the best two state-of-the-art models against our added contributions. The precision recall curves highlighting important model comparisons can also be seen in Appendix B. Furthermore, we carry out paired t-tests to showcase the statistical significance of improvements.

Table 2: Anomaly detection results on whole-body PET hold-out testing data. The performance is measured with best achievable DICE-score ($\lceil DICE \rceil$) and AUPRC on the test set.

Method	Anomalous Training Data		Normal Training Data	
	$\lceil DICE \rceil$	AUPRC	$\lceil DICE \rceil$	AUPRC
AE (Dense) (Baur et al., 2020)	0.333	0.301	0.371	0.322
AE (Dense) + CT (Baur et al., 2020)	0.332	0.281	0.360	0.319
AE (Spatial) Baur2020	0.313	0.251	0.355	0.349
AE (Spatial) + CT (Baur et al., 2020)	0.354	0.315	0.377	0.325
AE (SSIM) (Baur et al., 2020)	0.349	0.315	0.355	0.352
AE (SSIM) + CT (Baur et al., 2020)	0.346	0.310	0.347	0.328
F-AnoGan (Schlegl et al., 2017)	0.366	0.361	0.401	0.384
VAE (Dense) (Baur et al., 2020)	0.371	0.381	0.422	0.392
VAE (Dense) + CT (Baur et al., 2020)	0.351	0.342	0.419	0.396
VQ-GAN + Transformer (Codebook optimised 3D GAN variant of (Pinaya et al., 2021))	0.463	0.410	0.509	0.453
VQ-GAN + Transformer + CT conditioning (ours)	0.497	0.473	0.551	0.521
VQ-GAN + Transformer + CT conditioning + KDE (ours)	0.562	**0.561**	0.575	**0.579**
VQ-GAN + Transformer + CT conditioning + KDE + 40% Thresholding (ours)	**0.598**	0.532	**0.612**	0.551

4.6.1 ABLATION STUDY:

For the models trained on anomalous training data we observe a statistically significant improvement ($P < 0.005$) in anomaly detection DICE performance by implementing CT conditioning compared to the 3D VQ-GAN variant approach of Pinaya et al. (2021). A statistically significant improvement in AUCPR is also recorded ($P < 0.001$). This result confirms our initial expectations on the use case of anatomical context in the case of whole-body PET. Given the variability of healthy radiotracer uptake patterns, it is expected that beyond common areas like the bladder, further context is required to identify uptake as physiological or pathological. By incorporating model uncertainty to generate KDE maps, we see a further improvement in the overall DICE score and an even greater increase in AUPRC from 0.473 to 0.561 against the CT conditioned model ($P < .001$). This behaviour can be explained by the increased variability around heterogeneous areas of healthy uptake,

attributing to a decrease in false positives. The main advantage, as visualised in Fig. 7, is the increase in precision. By discarding the assumption of Gaussian uptake distributions, the model can better differentiate patterns of physiological uptake from pathological whilst still being sensitive to subtle anomalies, as seen in sample C in Fig. 7. We also see the same improvements across models in DICE and AUPRC for the models trained on normal data.
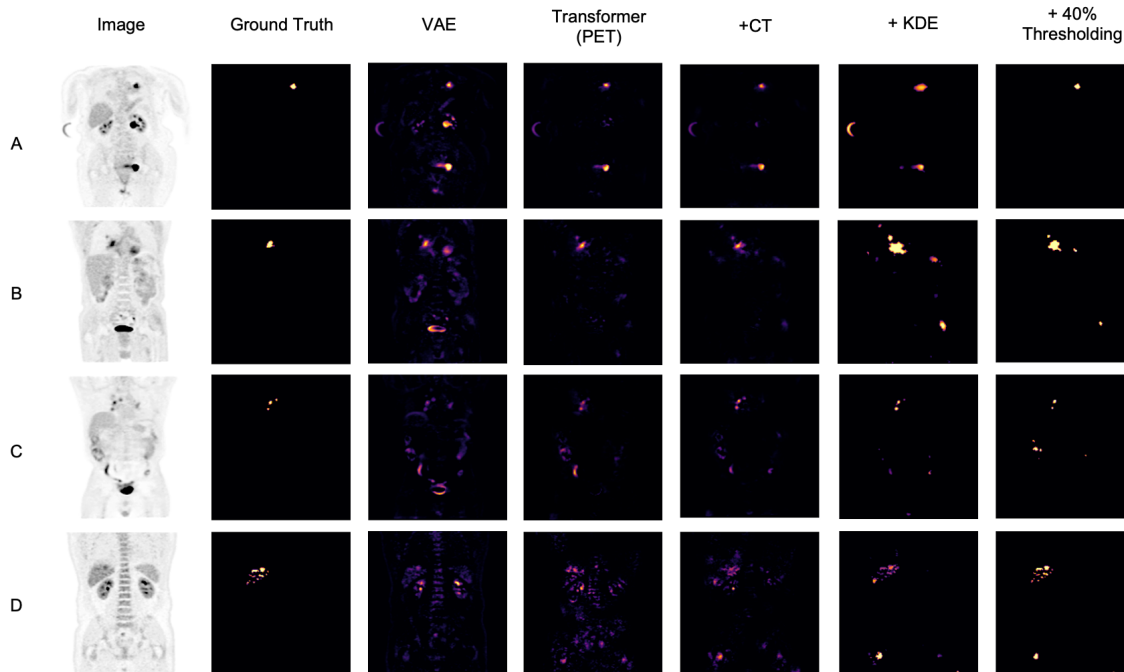
4.6.2 Comparison to state-of-the-art:

Figure 7: Columns from left to right display (1st) the input image; (2nd) the gold standard truth segmentation; (3rd) the anomaly map as the residual for the PET only VAE, (4th) Transformer, and (5th) CT conditioned methods; (6th) the abnormality map as a KDE, (7th) and after thresholding at 40% of each abnormal region maximum values. Results are provided for four randomly chosen subjects (A,B,C,D)

From Table 2, we can see a statistically-significant improvement ($P < .001$) presented via the VQ-GAN + transformer approach using only PET data in relation to all autoencoders with and withouth CT inputs, the variational autoencoder and F-AnoGan method. This result is expected, as demonstrated in prior research (Pinaya et al., 2021). However, this divergence is also attributed to the presence of anomalies during training. It can be observed in some samples that the autoencoder method performs worse on large anomalies as it is able to partially or fully reconstruct them. We can see looking at the autoencoder methods trained on healhty data that performance has improved, however even given this, our proposed method still outperforms all baselines. Comparing the method proposed by

Pinaya et al. (2021) to our best model comprising of CT conditioning and KDE anomaly maps, our approach generates an improvement in DICE score from 0.463 to 0.562 ($P < .001$) with a considerable increase in AUPRC from 0.410 to 0.561 ($P < .001$). Finally, through clinically accurate segmentations by growing segmented regions, we see a large increase in the best possible DICE score, but a reduction in AUPRC, likely brought about by the expansion of false-positive regions.

4.7 Golden Out-of-sample Testing

As a further test of the proposed models and their ability to generalize to unseen data from a different source/dataset, we showcase the performance of each proposed model and state-of-the-art comparisons on fully out-of-sample data using the CPTAC dataset. Pictorial results can be seen in Fig. 12, with numerical performance visible in table 3. Note that the performance on the golden hold-out sample was, surprisingly, found to be better than in the previous experiments, suggesting that Dice scores are highly dependent on the type and size of the cancers.

Table 3: Anomaly detection results on out-of-sample testing data. The performance is measured with best achievable DICE-score ($\lceil DICE \rceil$) and AUPRC on the test set.

Method	Anomalous Training Data		Normal Training Data	
	$\lceil DICE \rceil$	AUPRC	$\lceil DICE \rceil$	AUPRC
AE (Dense) (Baur et al., 2020)	0.359	0.344	0.438	0.386
AE (Dense) + CT (Baur et al., 2020)	0.437	0.403	0.423	0.374
AE (Spatial) Baur2020	0.351	0.338	0.382	0.348
AE (Spatial) + CT (Baur et al., 2020)	0.358	0.342	0.385	0.349
AE (SSIM) (Baur et al., 2020)	0.371	0.374	0.376	0.379
AE (SSIM) + CT (Baur et al., 2020)	0.365	0.373	0.364	0.371
F-AnoGan (Schlegl et al., 2017)	0.424	0.408	0.415	0.375
VAE (Dense) (Baur et al., 2020)	0.402	0.371	0.436	0.413
VAE (Dense) + CT (Baur et al., 2020)	0.394	0.369	0.447	0.408
VQ-GAN + Transformer (Codebook optimised 3D GAN variant of (Pinaya et al., 2021))	0.453	0.385	0.471	0 .405
VQ-GAN + Transformer + CT conditioning (ours)	0.500	0.468	0.529	0.511
VQ-GAN + Transformer + CT conditioning + KDE (ours)	0.610	0.604	0.618	0.600
VQ-GAN + Transformer + CT conditioning + KDE + 40% Thresholding (ours)	**0.717**	**0.631**	**0.719**	**0.642**

5. Discussion

From the results, there is strong evidence and motivation for using multi-modal conditioning for whole-body PET anomaly detection and using a KDE approach for producing anomaly maps. Not only do the proposed methods generate improved results over current state-of-the-art, but their performance is able to generalise to out-of-sample data and perform to the same level of competency. Firstly, we demonstrate the impact of appropriate codebook sizes when employing discrete latent space based representations for anomaly detection with autoregressive models. This alone showed a large range in performance based on changes using PET data only; additionally, this impact was further shown with respect to the size of the codebook for the conditioning CT data. Furthermore, with respect to the KDE approach for generating anomaly maps, we showcase the importance of an appropriate kernel shape and bandwidth regularisation term, which can generate large performance
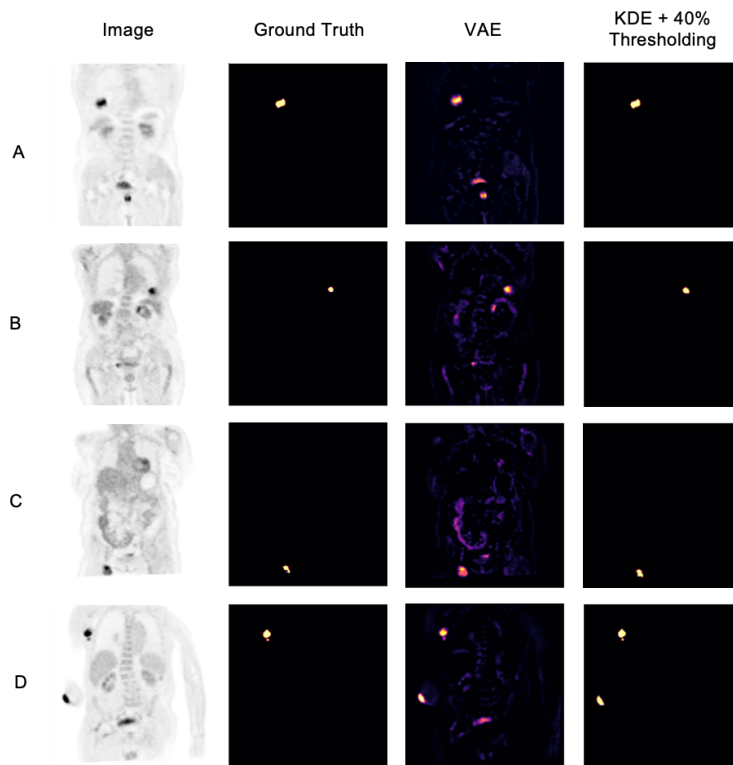
Figure 8: Columns from left to right display (1st) the input image; (2nd) the gold standard truth segmentation; (3rd) the anomaly map as the residual for the VAE, (4th) our approach using the abnormality map as a KDE and after thresholding at 40% of each abnormal region maximum values. Results are provided for four randomly chosen subjects (A,B,C,D)

improvements when optimised properly. We should note however that the efficacy of CT conditioning is greatly brought about through the attention mechanism used in transformer that allows the model to effectively model the relationship between the PET and CT data. We can see in the baseline methods that when both PET and CT data is used, there is little to no performance increase. Additionally we showcase how our approach works well on models trained on both normal and anomalous training data. We can see that there is a slight improvement when trained on normal data, as to be expected, however in the field of medical imaging where normal data can often be hard to come by, being able to produce a model via unsupervised approaches whilst still generating comparable results is a key breakthrough of this work. There are, however, still areas for improvement beyond the current scope of this research. We still see varying cases of false positives across samples, showing ongoing difficulties in differentiating physiological from pathological uptake. The reasons for this are likely patient-specific and can be down to several factors, i.e., pre-existing diseases, general health, age, PET/CT miss-alignment, or whether the scan was performed in fasting state or not. A further example can be seen in sample A, Fig 7,

where the injection site can be visualised in the patient's arm (although traditionally PET scans are performed in the "arms up" position). Additionally, we see in Sample D in Fig. 12 that a patient with an amputated arm showcases high uptake at the amputation site recorded as a false positive. Naturally, one solution would be to provide more training data increasing observed variability; however, further work to combat these issues could stem from some form of weak or self-supervision via to help to remove the outliers from the training data. Another approach could also be to incorporate some form of ensembling of models at different levels of downsampling within the VQ-GAN. In doing so the combined models will showcase greater local and global context such that the model might be able to better differentiate healthy physiological uptake from pathological uptake.

6. Conclusion

Detection and segmentation of anomalous regions, particularly for cancer patients, is essential for staging, treatment and intervention planning. In this study, we propose a novel transformer-based anomaly detection model using multi-modal conditioning and kernel density estimation via model stochasticity. Proposed model achieves statistically-significant improvements in Dice and AUPRC, representing a new state-of-the-art compared to competing methods. We further show the impact of codebook size selection to act as a key consideration when implementing VQ-VAE based methods. In addition, we show that the kernel choice and bandwidth regularisation for the kernel density estimation approach significantly impact the anomaly detection performance when using KDE anomaly maps, a superior alternative to residual maps. We show the impact of proposed methods when faced only with training data containing anomalies, showing greater robustness than autoencoder-only approaches. The strong evidence presented here indicates that multi-modal abnormality detection models, when combined with the proposed KDEs method, are key features that deserve further focus and development by the community.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

References

A S Ahmad, N Ormiston-Smith, and P D Sasieni. Trends in the lifetime risk of developing cancer in great britain: comparison of risk for those born from 1930 to 1960. *British journal of cancer*, 112:943–7, 3 2015. ISSN 1532-1827. doi: 10.1038/bjc.2014.606.

Varghese Alex, Mohammed Safwan K. P., Sai Saketh Chennamsetty, and Ganapathy Krishnamurthi. Generative adversarial networks for brain lesion detection. page 101330G, 2 2017. doi: 10.1117/12.2254487.

Ahmad Almuhaideb, Nikolaos Papathanasiou, and Jamshed Bomanji. 18f-fdg pet/ct imaging in oncology. *Annals of Saudi medicine*, 31:3–13. ISSN 0975-4466. doi: 10.4103/0256-4947.75771.

Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. 6 2018. doi: 10.1016/j.compmedimag.2019.101684.

Karim Armanious, Chenming Jiang, Sherif Abdulatif, Thomas Küstner, Sergios Gatidis, and Bin Yang. Unsupervised medical image translation using cycle-medgan. 3 2019. doi: 10.23919/EUSIPCO.2019.8902799.

Hans E. Atlason, Askell Love, Sigurdur Sigurdsson, Vilmundur Gudnason, and Lotta M. Ellingsen. Unsupervised brain lesion segmentation from mri using a convolutional autoencoder. 11 2018. doi: 10.1117/12.2512953.

Shaimaa Bakr, Olivie Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Weiruo Zhang, Ann Leung, Michael Kadoch, Joseph Shrager, Andrew Quon, Daniel Rubin, Sylvia Plevritis, and Sandy Napel. Data for nsclc radiogenomics collection, 2017.

Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann N. C. Leung, Michael Kadoch, Chuong D. Hoang, Joseph Shrager, Andrew Quon, Daniel L. Rubin, Sylvia K. Plevritis, and Sandy Napel. A radiogenomic dataset of non-small cell lung cancer. *Scientific Data*, 5:180202, 12 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.202.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images, 2019.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. 4 2020.

Avi Ben-Cohen, Eyal Klang, Stephen P. Raskin, Michal Marianne Amitai, and Hayit Greenspan. Virtual pet images from ct data using deep convolutional networks: Initial results. 7 2017. doi: 10.1007/978-3-319-68127-6_6.

Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *VISIGRAPP*, 2019.

Beatrice Berthon, Emiliano Spezi, Paulina Galavis, Tony Shepherd, Aditya Apte, Mathieu Hatt, Hadi Fayad, Elisabetta De Bernardi, Chiara D. Soffientini, C. Ross Schmidtlein, Issam El Naqa, Robert Jeraj, Wei Lu, Shiva Das, Habib Zaidi, Osama R. Mawlawi, Dimitris Visvikis, John A. Lee, and Assen S. Kirov. Toward a standard for the evaluation of pet -auto-segmentation methods following the recommendations of aapm task group no. 211: Requirements and implementation. *Medical Physics*, 44:4098–4111, 8 2017. ISSN 0094-2405. doi: 10.1002/mp.12312.

Lei Bi, Jinman Kim, Ashnil Kumar, Dagan Feng, and Michael Fulham. Synthesis of positron emission tomography (pet) images via multi-channel generative adversarial networks (gans). 7 2017.

Ninon Burgos, M Jorge Cardoso, Jorge Samper-González, Marie-Odile Habert, Stanley Durrleman, Sébastien Ourselin, , , Olivier Colliot, Alzheimer's Disease Neuroimaging Initiative, and Frontotemporal Lobar Degeneration Neuroimaging Initiative. Anomaly detection for the individual analysis of brain pet images. *Journal of medical imaging (Bellingham, Wash.)*, 8:024003, 3 2021. ISSN 2329-4302. doi: 10.1117/1.JMI.8.2.024003.

Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. 3 2021.

Mark Chen, Alec Radford, Jeff Wu, Jun Heewoo, and Prafulla Dhariwal. Generative pre-training from pixels. 2020.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. 11 2016.

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. 6 2018.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. 4 2019.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. 9 2020.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26:1045–1057, 12 2013a. ISSN 0897-1889. doi: 10.1007/s10278-013-9622-7.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and

Fred Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26:1045–1057, 12 2013b. ISSN 0897-1889. doi: 10.1007/s10278-013-9622-7.

CPTAC. National cancer institute clinical proteomic tumor analysis consortium, 2018.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. 4 2020.

Alexander Drzezga, Timo Grimmer, Matthias Riemenschneider, Nicola Lautenschlager, Hartwig Siebner, Panagiotis Alexopoulus, Satoshi Minoshima, Markus Schwaiger, and Alexander Kurz. Prediction of individual clinical outcome in mci by means of genetic assessment and (18)f-fdg pet. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 46:1625–32, 10 2005. ISSN 0161-5505.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. 6 2016.

Keigo Endo, Noboru Oriuchi, Tetsuya Higuchi, Yasuhiko Iida, Hirofumi Hanaoka, Mitsuyuki Miyakubo, Tomohiro Ishikita, and Keiko Koyama. Pet and pet/ct using 18f-fdg in the diagnosis and management of cancer patients. *International Journal of Clinical Oncology*, 11:286–296, 9 2006. ISSN 1341-9625. doi: 10.1007/s10147-006-0595-0.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. 12 2020.

Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9:601, 10 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01718-3.

Olivier Gevaert, Jiajing Xu, Chuong D. Hoang, Ann N. Leung, Yue Xu, Andrew Quon, Daniel L. Rubin, Sandy Napel, and Sylvia K. Plevritis. Non–small cell lung cancer: Identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results. *Radiology*, 264:387–396, 8 2012. ISSN 0033-8419. doi: 10.1148/radiol.12111607.

Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-attention is all you need: Adapting pretrained transformers for machine translation. 4 2021.

Mathieu Hatt, John A. Lee, Charles R. Schmidtlein, Issam El Naqa, Curtis Caldwell, Elisabetta De Bernardi, Wei Lu, Shiva Das, Xavier Geets, Vincent Gregoire, Robert Jeraj, Michael P. MacManus, Osama R. Mawlawi, Ursula Nestle, Andrei B. Pugachev, Heiko Schöder, Tony Shepherd, Emiliano Spezi, Dimitris Visvikis, Habib Zaidi, and Assen S. Kirov. Classification and evaluation strategies of auto-segmentation approaches for pet: Report of aapm task group no. 211. *Medical Physics*, 44:e1–e42, 6 2017. ISSN 00942405. doi: 10.1002/mp.12124.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. pages 5967–5976. IEEE, 7 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.632.

Eunhee Kang, Hyun Jung Koo, Dong Hyun Yang, Joon Bum Seo, and Jong Chul Ye. Cycle consistent adversarial denoising network for multiphase coronary ct angiography. 6 2018. doi: 10.1002/mp.13284.

Hyun Su Kim, Kyung Soo Lee, Yoshiharu Ohno, Edwin J R van Beek, and Juergen Biederer. Pet/ct versus mri for diagnosis, staging, and follow-up of lung cancer. *Journal of magnetic resonance imaging : JMRI*, 42:247–60, 8 2015. ISSN 1522-2586. doi: 10.1002/jmri.24776.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 12 2014.

Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. 5 2017.

Bin Liu, Sujuan Gao, and Shuofeng Li. A comprehensive comparison of ct, mri, positron emission tomography or positron emission tomography/ct, and diffusion weighted imaging-mri for detecting the lymph nodes metastases in patients with cervical cancer: A meta-analysis based on 67 studies. *Gynecologic and Obstetric Investigation*, 82:209–222, 2017. ISSN 0378-7346. doi: 10.1159/000456006.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. 11 2015.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. pages 2813–2821. IEEE, 10 2017. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.304.

Sergio Naval Marimont and Giacomo Tarroni. Anomaly detection through latent space restoration using vector-quantized variational autoencoders. 12 2020.

Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. pages 416–425. IEEE, 6 2020. ISBN 978-1-7281-9360-1. doi: 10.1109/CVPRW50498.2020.00054.

David E Newman-Toker, Zheyu Wang, Yuxin Zhu, Najlla Nassery, Ali S Saber Tehrani, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Mehdi Fanai, and Dana Siegal. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "big three". *Diagnosis (Berlin, Germany)*, 8:67–84, 2021. ISSN 2194-802X. doi: 10.1515/dx-2019-0104.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076, 9 1962. ISSN 0003-4851. doi: 10.1214/aoms/1177704472.

Ashay Patel, Petru-Daniel Tudosiu, Walter Hugo Lopez Pinaya, Gary Cook, Vicky Goh, Sebastien Ourselin, and M Jorge Cardoso. Cross attention transformers for multi-modal unsupervised whole-body pet anomaly detection. In *MICCAI Workshop on Deep Generative Models*, pages 14–23. Springer, 2022.

Daniela Perani, Perani Daniela, Orazio Schillaci, Schillaci Orazio, Alessandro Padovani, Padovani Alessandro, Flavio Mariano Nobili, Nobili Flavio Mariano, Leonardo Iaccarino, Iaccarino Leonardo, Pasquale Anthony Della Rosa, Della Rosa Pasquale Anthony, Giovanni Frisoni, Frisoni Giovanni, Carlo Caltagirone, and Caltagirone Carlo. A survey of fdg- and amyloid-pet imaging in dementia and grade analysis. *BioMed research international*, 2014:785039, 2014. ISSN 2314-6141. doi: 10.1155/2014/785039.

Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Unsupervised brain anomaly detection and segmentation with transformers. 2 2021.

Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. Compressed sensing mri reconstruction using a generative adversarial network with a cyclic loss. *IEEE Transactions on Medical Imaging*, 37:1488–1497, 6 2018. ISSN 0278-0062. doi: 10.1109/TMI.2018.2820120.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

Dimitri Renard, Rik Vandenberghe, Laurent Collombier, Pierre-Olivier Kotzki, Jean-Pierre Pouget, and Vincent Boudousq. Glucose metabolism in nine patients with probable sporadic creutzfeldt-jakob disease: Fdg-pet study using spm and individual patient analysis. *Journal of neurology*, 260:3055–64, 12 2013. ISSN 1432-1459. doi: 10.1007/s00415-013-7117-6.

Daisuke Sato, Shouhei Hanaoka, Yukihiro Nomura, Tomomi Takenaga, Soichiro Miki, Takeharu Yoshikawa, Naoto Hayashi, and Osamu Abe. A primitive study on unsupervised anomaly detection with an autoencoder in emergency head ct volumes. page 60. SPIE, 2 2018. ISBN 9781510616394. doi: 10.1117/12.2292276.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. 3 2017.

Scikit-Learn. Density estimation. URL https://scikit-learn.org/stable/modules/density.html.

M Signorini, E Paulesu, K Friston, D Perani, A Colleluori, G Lucignani, F Grassi, V Bettinardi, R S Frackowiak, and F Fazio. Rapid assessment of regional cerebral metabolic abnormalities in single subjects with quantitative and nonquantitative [18f]fdg pet: A clinical validation of statistical parametric mapping. *NeuroImage*, 9:63–80, 1 1999. ISSN 1053-8119. doi: 10.1006/nimg.1998.0381.

B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, 2 2018. ISBN 9781315140919. doi: 10.1201/9781315140919.

Liyan Sun, Jiexiang Wang, Yue Huang, Xinghao Ding, Hayit Greenspan, and John Paisley. An adversarial learning approach to medical image synthesis for lesion detection. 10 2018.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71:209–249, 2021. ISSN 1542-4863. doi: 10.3322/caac.21660.

Shinji Takaki, Toru Nakashika, Xin Wang, and Junichi Yamagishi. Stft spectral loss for training a neural speech waveform model. 10 2018.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. 11 2020.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. 11 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 6 2017.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. 4 2021.

Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019. ISSN 1361-8423. doi: 10.1016/j.media.2019.101552.

Ke Zhang and Xiahai Zhuang. Shapepu: A new pu learning framework regularized by global consistency for scribble supervised cardiac segmentation, 2022.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. 1 2018.

Maria A. Zuluaga, Don Hush, Edgar J. F. Delgado Leyton, Marcela Hernández Hoyos, and Maciej Orkisz. Learning from only positive and unlabeled data to detect lesions in vascular ct images, 2011.

Appendix A. Model Parameters

The total number of parameters for the VQ-GAN Model used to encode the PET and CT along with the parameters of the Performer models used in this work can be outlined in Table 4.

Table 4: Number of parameters of models used

Model	Trainable Parameters	Total Parameters
VQ-GAN	24,599,681	24,632,449
Performer PET only	18,684,673	18,684,673
Performer PET + Cross attention CT	33,396,993	33,396,993

Appendix B. Model AUC Curves

Furthermore we display a number of meaningful Precision Recall curves for selected models on both testing datasets to better visualise the sensitivity and specificity tradeoff of the proposed methods.
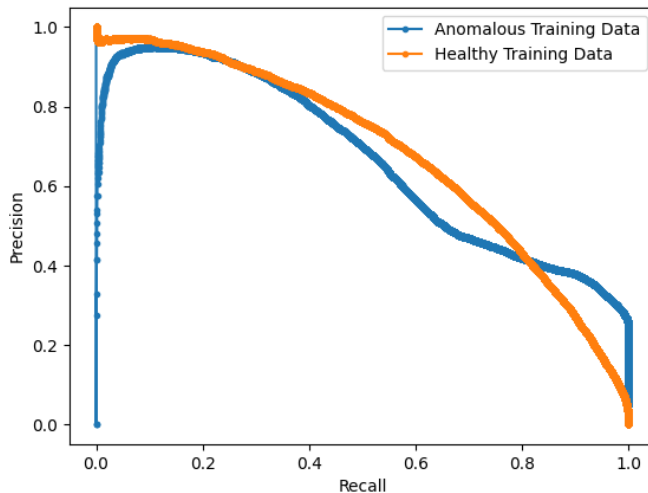

Figure 9: Precision recall curves for proposed model using PET and CT data with KDE anomaly maps trained on anomalous and healthy training data, for testing outlined in section 4.6
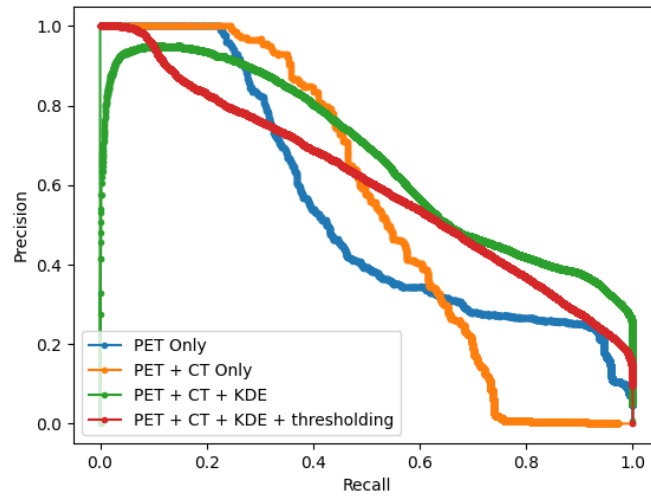
Figure 10: Precision recall curves for ablation study of VQ-VAE + Transformer along with additions including CT conditioning, KDE anomaly maps and thresholding. Results presented for testing outlined in section 4.6
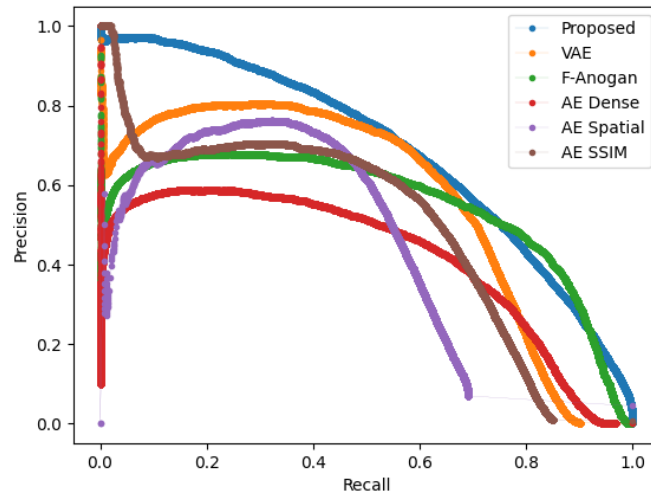

Figure 11: Precision recall curves comparing baseline methods against our proposed method of the VQ-VAE + Transformer with CT conditioning and KDE anomaly map trained on healthy data. Results presented for testing outlined in section 4.6
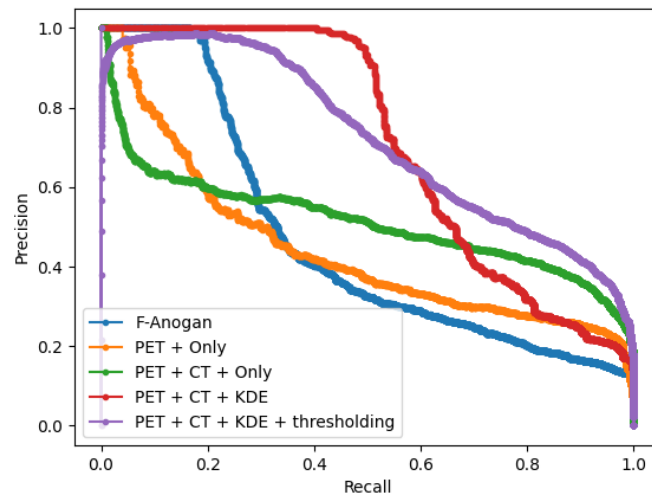
Figure 12: Precision recall curves for ablation study of VQ-VAE + Transformer along with additions including CT conditioning, KDE anomaly maps and thresholding. Additionally we showcase the performance of the best alternative baseline model for the CPTAC testing set - F-AnoGan. Results presented for testing on the out-of-sample testing on CPTAC for models trained on anomlaous data.